Bioinformatics Summit Meeting Report

May 8-9, 2007

This Bioinformatics Summit was the second Bioinformatics Summit organized by the Division of Allergy, Immunology and Transplantation (DAIT). It was held on May 8-9, 2007 at Courtyard Marriott in Gaithersburg, Maryland. There were 110 people registered for the meeting and 64 people attended the meeting. The participants were from many NIH institutes including the National Institute of Allergy and Infectious Diseases (NIAID), the National Cancer Institute (NCI), the National Institute on Aging (NIA), the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), the National Institute of General Medical Sciences (NIGMS), the National Institute on Deafness and Other Communication Disorders (NIDCD), the National Library of Medicine (NLM), the National Institute on Alcohol Abuse and Alcoholism (NIAAA), the Center for Information Technology (CIT), National Center for Complementary and Alternative Medicine (NCCAM), and other Federal agencies such as the National Institute of Standards and Technology (NIST) and the National Science Foundation (NSF). Additionally, non-governmental organizations such as Institute for Systems Biology, MITRE, Juvenile Diabetes Research Foundation, University of Maryland, Columbia University, Northrop Grumman, SRA, Georgetown University and DAIT funded grantees and contractors attended.

In this meeting, we successfully reached our goals to:

- review ongoing work on immunology related ontologies and data standards
- review ongoing work in patient privacy when sharing data in bioinformatics
- identify synergies in Bioinformatics among DAIT and other divisions of NIAID, i.e. the Division of Acquired Immunodeficiency Syndrome (DAIDS) and the Division of Microbiology and Infectious Diseases (DMID), and other government agencies such as FDA
- learn from ongoing projects such as caBIG and from other knowledge management systems such as Wiki Professional

Dr. Mark Musen chaired the session on ontologies and data standards in DAIT-funded projects. Dr. Alexander D. Diehl, senior scientific curator at Mouse Genome Informatics of The Jackson Laboratory introduced his collaboration with the Bioinformatics Integration Support Contract (BISC) team and others in enhancing the Gene Ontology (GO) to expand GO's coverage in immunology. This effort and continued gene annotation using the GO immunology will greatly enhance our understanding of immunology related genes among different model organisms and lay a foundation for using the GO immunology in computational biology and systems biology.

Dr. Ryan Brinkman, senior scientist at Terry Fox Laboratory BC Cancer Research Centre and assistant professor at Medical Genetics of University British Columbia, presented data standards for flow cytometry. Leading a group of scientists composed of both flow cytometry vendors and users including the BISC team, the Immune Tolerance Network (ITN) and other DAIT funded principle investigators, he and others working together developed a set of minimum data elements to capture flow cytometry data. This is a very useful minimal information set to enhance data interoperability in exchanging and central archiving of flow cytometry data.

Dr. Lindsay Cowell, assistant professor at Department of Biostatistics and Bioinformatics Duke University, Laboratory of Computational Immunology Center for Bioinformatics and Computational Biology, introduced her work in representation of complex immunological networks through the use of ontologies funded by the DAIT immune modeling center contract. She emphasized the importance of developing a new ontology in the context of the OBO Foundry and presented her preliminary results in developing new relationships to model immune processes in complex immunological networks.

Dr. Mark Musen from Stanford University highlighted new progress in ontology development at the National Center for Biomedical Ontology. He introduced the new web site Bioportal (http://www.bioontology.org) launched in February 2007. This web site presents a new tool for browsing and searching terms across ontologies, which is an excellent tool for disseminating ontologies and facilitating the development of new ontologies.

Dr. Effie Petersdorf, chair of the HLA Genetics Steering Committee at Fred Hutchinson Cancer Research Center, gave an overview of HLA terminology standards and challenges in biomedical research. She introduced the changes in the history of evolving HLA identification technologies and versions of HLA nomenclatures. She presented the hierarchical structures among HLA terms used for different resolutions of HLA typing techniques and the divergence of historical terms from current terms. The lack of correct semantic mapping in HLA terms poses great challenges for data interoperability among databases.

The following issues were raised during the discussion in this session.

- Training on the development of ontology and how to use ontology are urgently needed. Although the list of ontologies on the OBO Foundry is growing rapidly, we are still short of good ontologies. Many ontologies were developed because of the needs of researchers, however many of these ontologies are plagued with problems. Many areas of biology still do not have good ontologies with which to work. Training of scientists by experienced ontologists is necessary to create high quality domain specific ontologies.
- Ontology evaluation criteria are important aspects of ontology development. Real use cases for ontology should be one of the crucial elements among many other evaluation criteria.
- Software tools should be developed to enhance the utility of ontologies such as a tool to extract terms from one ontology and merge them into another ontology semantically. A tool to deploy or configure applications with ontology is also needed, such as a tool to deploy clinical trial protocol ontology to configure a clinical trial data management system dynamically.

Dr. David Karp, University of Texas Southwestern Medical Center, chaired the second session, challenges in data sharing and data reusability. Dr. Stephen E. Wilson, Director of the Division of Biometrics III from the Center for Drug Evaluation and Research (CDER) at the US Food & Drug Administration (FDA), gave a presentation entitled ‘FDA's Critical Path and CDER's Office of Translational Sciences: Opportunities for Collaboration.’ He highlighted the potential bioinformatics synergies among the FDA, industry and NIH for building infrastructures to support new product development, sharing existing knowledge and databases, and developing data standards.

Dr. Bradley Malin, Assistant Professor of Biomedical Informatics at Vanderbilt University Medical Center and Assistant Professor of Computer Science at Vanderbilt University laid out the challenges protecting patient privacy when sharing research data imposed by information technology and public accessible data. He presented computational approaches and the limitations of these approaches for patient privacy protection in secondary data sharing when faced with the fact that 87% of the United States population is re-identifiable by combining two publicly accessible datasets.

Dr. Karp gave an overview of the BISC bioethics meeting held at University Texas South Western Medical School on Jan. 17, 2007. He reviewed eight recommendations from an expert panel of bioethicists, scientists, legal experts and IRB representatives about data submission and data dissemination through the Immport, a bioinformatics application of the BISC project. These recommendations are:

1. Confirm ethics oversight of initial data collection by the local IRB and investigator.
2. Establish a mechanism to review the use of data within Immport.
3. Going forward, informed consent should take into account the incorporation of data into shared databases such as ImmPort.
4. The use of data obtained from legacy databases creates special challenges for patient consent. This group suggested having a policy in place to deal with potential problems with using legacy data.
5. Pursue efforts directed at standardization of data.
6. Establish data sharing rules, including attribution of contributions.
7. Adopt "best practices" to avoid the possibility of identifying study participants from clinical data.
8. Periodically review and publicize the ethical guidelines used by the BISC project in ImmPort.

Dr. Anita L. DeStefano, Co-Director Biostatistics Program and Associate Professor of Biostatistics and Neurology at Boston University Schools of Public Health and Medicine presented her experience in managing data generated from large-scale genetic studies in the Framingham Heart Study. This is one of the datasets that can be accessed through dbGAP, a public database hosted by NCBI. She introduced their practice of managing patient consent on an individual participant level. The research data generated from 9500 participants in the Framingham Heart Study were disseminated based on the consent terms agreed to by individual participants.

Issues discussed in the second session discussion period included:

- the best ways to enforce data standards to facilitate data sharing among the scientific community, industry and government agencies
- the need to separate sharing biological specimens and sharing research data in the patient consent form
- HIPAA compliance is not secure enough to protect the identity of the research participant. However, committee-review based data access to genetic and clinical research data may create privileged classes of users, which may not be a fair practice

Associate Director of the Office of Biomedical Informatics at DAIT, Ms. Cheryl Kraft, chaired the third session, Data Interoperability in NIAID Funded Projects. Dr. Richard Scheuermann gave an update on the BISC project and data interoperability. Data interoperability is the biggest challenge for the BISC project. Dr. Scheuermann envisioned data interoperability as a set of minimum data standards plus standardized ontologies and an extensible data model. In collaboration with other DAIT-funded programs and other scientific ontology communities, the BISC team has developed minimum information set standards and standardized ontologies. The BISC team also developed a generic data model that meets the diverse needs of research data generated by the immunology research community.

Executive Director of Bioinformatics at the Immune Tolerance Network (ITN), Mr. David Parrish, presented a clinical trial informatics system developed at the ITN. At the ITN, clinical study activities were converted into data structures using a structured and extendible ontology based on the elements and relationships required to describe a protocol of a clinical study. The same ontology was also used to configure the data collection, data storage and the interface for data transformation. On top of this ontology-based data repository, a visualization tool, ‘Dashboard’, was implemented to support user-defined data reports and data graphics.

Ms. Valentina Di Francesco, program officer from the Division of Infectious Diseases (DMID), gave an overview of bioinformatics resources at DMID. She introduced their effort to make data interoperable among eight groups funded through the Bioinformatics Research Centers contract (BRC). She also discussed their effort to evaluate public awareness of bioinformatics in the scientific community. This evaluation was done in order to tailor the bioinformatics resources within the BRCs to match the level of bioinformatics capabilities of sceintific researchers in the their research community.

Dr. Jonathan Kagan, Deputy Director of Program Development at the Division of AIDS (DAIDS), introduced the DAIDS Enterprise System. He emphasized how the DAIDS Enterprise System revolutionized the way DAIDS program officers monitoring clinical studies of AIDS treatment on a global scale performed their work and enhanced the speed and the accuracy of communications within DAIDS and in NIAID.

Discussions in this session were focused on:

- the possibility of sharing clinical trial results with other research investigators
- the management of human factors in data sharing and data interoperability,
- the use of technology to enhance and enforce the use of common vocabularies when capturing protocol information, adverse event reporting and clinical trial data analysis.

Dr. Richard Scheuermann chaired the last session of the summit: Technology Demonstration. Dr. Douglas Fridsma, Assistant Professor of Medicine at University of Pittsburgh presented lessons learned in the BRIDG project of caBIG. He emphasized following points that would be very useful to guide the data interoperability effort at DAIT:

- Start from a use-case driven scope to solve a problem that exists
- Capture knowledge early in the development cycle by focusing on the requirements from domain experts
- Keep the model free of implementation-specific formalisms but as practical as possible
- Imbed the model development in a process that is scaleable and reproducible
- Make the data model collaboratively agreeable in the community and standards-based

Dr. Barend Mons, Associate Professor in the Department of Medical Informatics**,** Leiden, University Medical Center Rotterdam, gave a demonstration of a new approach to barcode the knowledge implemented on the WikiProfessional web. He introduced a computer-base knowledge extraction system using terms and clusters of terms to form "knowlets" of scientific concepts. The relationships among the knowlets were established by an algorithm built upon existing ontology and the frequency of co-occurrences of the terms and concepts. This trainable algorithm is used to scan through abstracts and scientific publications. These publications were tagged with a barcode of knowlets. The barcode was then used to facilitate the building of a knowledge base and was used to assist in query retrieval of related papers. The WikiProfessional pages that were edited by domain experts from the scientific community were also tagged with barcodes of knowlets. This knowlet extraction and barcode tagging algorithm can be updated dynamically according to advances in science presented in publications. Dr. Mons indicated that his algorithm can be trained with various ontologies and perform better than rigid ontologies.

Drs. Frisdma and Mons' presentations sparked discussions about data interoperability and data integration in basic and clinical research, ontology and ontology alternative technology in semantic mapping of research data. Additionally, the million mind approach of Wiki and its impact in the future of science was discussed.

The bioinformatics summit 2007 concluded with agreement that there needs to be more collaboration and better communication within the community. There is also a need for continued work on more focused, real use-case oriented projects that will support the immunology research community with better bioinformatics tools and standards.